R E P O R T   R E S U M E S

ED 012 923 AL 000 639

COMPUTATIONAL LINGUISTICS--PROCEDURES AND PROBLEMS.

BY- LEHMANN, W.P.

TEXAS UNIV., AUSTIN, LINGUISTICS RES. CTR.

REPORT NUMBER LRC-65-WA-1 PUB DATE JAN 65

EDRS PRICE MF-$0.25 HC-$1.48 37P.

DESCRIPTORS- *COMPUTATIONAL LINGUISTICS, LANGUAGE, LINGUISTIC THEORY, *MACHINE TRANSLATION, MATHEMATICAL LINGUISTICS, LINGUISTIC PATTERNS, STRUCTURAL ANALYSIS, *DATA PROCESSING, COMPUTERS, CLASSIFICATION,

BASED ON A LECTURE GIVEN AT THE UNIV. OF TEXAS SCIENCE CONFERENCE, NOV. 20, 1964, THIS PAPER PRESENTS IN RELATIVELY NON-TECHNICAL TERMINOLOGY A DESCRIPTION OF THE "STRUCTURAL" APPROACH TO THE STUDY OF LANGUAGE WHICH UNDERLIES THE WORK OF THE LINGUISTICS RESEARCH CENTER. THIS APPROACH ANALYZES LANGUAGE IN SUCH A WAY THAT IT CAN BE MANIPULATED WITH A COMPUTER. STRESSING THE NECESSITY FOR A MORE COMPLETE UNDERSTANDING OF LANGUAGE AS THE BASIS FOR MACHINE TRANSLATION AND COMPUTATIONAL LINGUISTICS, THE AUTHOR DEALS WITH (1) THE FORMAL STRUCTURE OF LANGUAGE, (2) SIMULATION, (3) LANGUAGE DATA PROCESSING, (4) AUTOMATIC CLASSIFICATION, (5) ANALYSIS OF MEANING, AND (6) ACCOMPLISHMENTS IN THE FIELD OF LINGUISTIC RESEARCH. INCLUDED ARE REPRODUCTIONS OF THE ANALYSIS OF A SENTENCE WITH A PARSING DIAGRAM, AND A CHART OF THE LINGUISTICS RESEARCH SYSTEM. (AM)

lrc

THIS IS A WORKING PAPER ☐ IT MAY BE EXPANDED. MODIFIED OR WITHDRAWN AT ANY TIME ☐ THE VIEWS, CONCLUSIONS, AND RECOMMENDATIONS EXPRESSED HEREIN DO NOT NECESSARILY REFLECT THE OFFICIAL VIEWS OF THE SPONSOR ☐

LINGUISTICS RESEARCH CENTER

THE UNIVERSITY OF TEXAS

BOX 7247 • UNIVERSITY STATION • AUSTIN 12, TEXAS

COMPUTATIONAL LINGUISTICS: PROCEDURES AND PROBLEMS

W. P. Lehmann

prepared for
National Science Foundation
Grant NSF GN-308

LINGUISTICS RESEARCH CENTER
The University of Texas
Box 7247, University Station
Austin, Texas 78712

LRC 65 WA-1

January 1965

CONTENTS

ABSTRACT

The necessity for a more complete understanding of language as the basis for machine translation and computational linguistics is stressed. Other benefits which will result from this longterm research -- including information retrieval and automatic classification -- are also mentioned.

FOREWORD

This paper is based on a lecture given at The University of Texas Science Conference, November 20, 1964. The conference provided a means for scientists in The University of Texas faculties to get acquainted with one another and to listen to brief expositions of research in progress on the campus. Part of the research mentioned here was performed under grant NSF GN-308.

1 INTRODUCTION

This paper deals with a relatively new approach to the study of language, one underlying the work of the Linguistics Research Center. This view regards language in such a way that it can be manipulated with a computer. Yet the view cannot be related to technological developments, for it preceded the computer. The resultant approach to language has often been called structural linguistics.

Language may be studied from many points of view. One may wish to acquire a graceful mastery of one or more languages, either for writing special kinds of texts called poems, or simply to impress as well as inform an audience. One may wish to learn about the history of specific languages, how English is related to Hindi, Greek, Armenian or Irish. The most prominent interest in language in Western culture arose from a desire to understand venerated texts, primarily texts in Hebrew, Greek, Latin, the Bible and the classics. The understanding of these texts led to the development of special techniques and attitudes about language. For us, oddly enough in contrast with the Greeks and Romans, the written language has seemed more fundamental than the spoken, and we have spent more time learning to read than to speak languages, whether French, German, Russian, or the cited classical languages. Further, since in our day written materials are broken up into units called words, these seem to us the fundamental entities of language.

Moreover, since these languages, especially Latin seem somehow to be model languages, we have sought a mastery of current languages, including our own, through descriptions--grammars--which are modeled on the grammar of Latin. To master a language, including English, our grammarians, teachers and students note its resemblances to Latin and also the differences from it. This procedure may be compared with that of a geographer who adopts one location, for example, New York City, as the ideal and describes all other locations by their resemblances to it. From such a geographer we would not get a map of London, Paris or Moscow, but rather various maps of New York modified in accordance with deviations from New York in these cities.

It is not my aim to present a critique of any view of language, or of our methods of teaching languages, or even of any type of research on language. But since we have all studied languages in accordance with the Latin-based approach, we regard any language in accordance with the views given us in our schools. These views must therefore be specified if we are to understand one another. I might also mention that the first attempts at computer processing of language failed because the scholars concerned viewed the essential problem as the manipulation of words.

Besides discussing a somewhat different approach to language I will touch on the linguistic investigations it has prompted and is continuing to require. I will also deal briefly with the requirements this approach is making on computer programming and may make on computer technology.

2 FORMAL STRUCTURE

Possibly the most important feature of structural linguistics is the understanding that language has a formal structure, composed of various sub-structures. In these structures the function or value of any entity is determined largely by its relationship with other entities. Entities then are not defined by their relationship to the outside world; a noun for example is not defined as the name of a person, place or thing. Viewing language in this way seems to a linguist somewhat similar to presenting mathematics through concrete objects, never to add 2 and 2, but always two apples to two apples, and so on. There is little doubt that our understanding of numbers, our progress in mathematics, would have been hampered if we had dealt with them only in connection with the outside world rather than as abstract signs. Linguists hold that such a view of language has impeded our understanding of it.

Some decades ago a few scholars began to examine language as a system of signs whose function was specified by their interrelationships. This approach to language--this theory, if you wish--would define a noun in any given language by its relationship to other entities in that language. A noun in English, for example, might be defined as an entity with certain relationships to inflectional elements, to a Z-like entity in the plural: arm : arms, or in the possessive: man : man's. Other languages might not have nominal inflection and accordingly would not have a class of nouns. In Japanese, for example, only verbs are inflected; we cannot then speak of a class of inflected nouns. Another basis of definition of entities might be by their relationship to independent entities, for example, articles: _the_, _a_, or to verbs. By such a definition _man_ is a noun because it can follow _the_; _went_ on the other hand is not. Using such a procedure, we can identify nouns in Japanese. Further, we can also identify larger acceptable entities, such as sentences, by their entities and the interrelationships of these; _men talk_, for example, is an acceptable English sentence, but not _men happily_, or even _men happy_.

When this approach to language was pursued, the work of linguistics came to be looked on as the determination of the entities of any language and their interrelationships.

Two requirements are necessary before one can deal usefully with language in this way. We must first determine whether the materials are genuine, whether for example an English speaker permits the sequence: _men talk_. Next, we ascertain whether the entities in this sequence have a characteristic meaning. In such determination we elicit comparable sequences, e.g. _men walk_, _then talk_, and so on. With such contrasting sequences we would satisfy ourselves that _m_ is a characteristic marker, for it is the only entity distinguishing _men talk_ from _then talk_, or distinguishing _man talks_ (a statement an anthropologist might make) from _Ann talks_ (a statement one might make about a very young lady). Similarly, _t_ is a characteristic sound marker, distinguishing _ten_ from _men_, _talk_ from _walk_, and so on.

In addition to determining the characteristic entities of sound in a language, we may also determine larger entities, for example, talk as opposed to walk. These differ from entities like the first consonants of men, ten and then in that they have established relationships to certain concepts. Briefly, we say they have meaning. The first consonants of men, ten and then do not. They serve to distinguish meanings, but we cannot associate with them any given concepts, such as 'animateness', 'number' or 'temporality'.

Rigorous techniques for determining entities of both kinds have been developed.

When such entities are specified in a given language, linguists set out to determine their role--one might say, their properties. In English there are about forty entities like m and t. These may be regarded as signs, comparable to other kinds of signs man uses, e.g. 3 4. Just as a mathematician might investigate relationships between such units in a given number system, setting up various classes, e.g. primes, so a linguist might investigate the role of such entities in a given language. He might determine what relationships t has with regard to the other entities of sound. In English, for example, t may precede e if n follows, but not e alone; there is no English sequence te. Nor are there sequences like tne, etn, and so on. Other such problems will occur to any linguist, mathematician, or to anyone who enjoys manipulating signs. Yet few such problems have been investigated, even in a widely used language like English, to say nothing of 5,000 other languages. We have not had the personal nor the physical resources.

A similar range of problems might be cited in the investigation of entities like walk talk take brake and so on. We might find sequences in which man: men precede any of these, e.g. the man walked, the man braked around curves, etc. But we do not find sequences like: walkman, talkman, takeman paralleling brakeman. A complete description of any language would specify which of such sequences occur.

At this stage of his investigations a linguist does not deal with meaning. He has determined that brake differs from take, that it has a meaning; but in examining possible sequences like brakeman he deals only with its properties of occurrence. Nonetheless this second type of investigation, noting the interrelationships between entities like take, brake, man, -ing, provides even more problems than does the first.

Still other entities must be identified in language and investigated similarly. But the two types of entities I have selected may exemplify the approach of structural linguistics.

Those structural linguists who concern themselves exclusively with the study of sets of linguistic entities and their interrelationships are sometimes called mathematical or computational linguists. Other linguists may deal with other language problems--the pronunciation of talk, walk in various areas, the stylistic differences between talk and speak, and so on. But a computational linguist limits his concern to sets of entities and their interrelationships.

If his approach to language is valid, in using language men acquire a number of entities and learn how to manipulate them in relation to one another. Further, if a machine could be devised which would store the number of entities stored by man, with rules specifying their relationships to other entities, the machine might simulate man's mastery of language.

3 SIMULATION

As is well-known, about twenty years ago a machine was developed which seemed to have the essential capabilities, the computer. Possibly the manipulation of language would never have engaged the attention of computer specialists if the problem of rapid intercommunication had not become so prominent. To be sure computation centers might have found it amusing in time to have a few language games available for visitors, when they became bored with tic-tac-toe, checkers, chess or go. But since the scientists, who were nursing along the infant computer and contemplating uses for it when it matured, had just been involved in international struggles which pointed up the importance of reading the scientific publications of the other side, they suggested that the computer might solve the problem of intercommunication. The computer therefore was looked on as the machine to take over the uninspiring activity of translation; supporting agencies provided time on computers and a small amount of money to research workers, whose goal was to be machine translation. This seemingly overriding goal was the prime activity for which language specialists might use computers. To the outside world today, still, linguists doing research with computers are working on machine translation.

With a million words a day of important materials awaiting translation from Russian to English, let alone materials of secondary importance or materials in Chinese, Japanese, German, French and so on, machine translation would be a fine accomplishment. But the problem requires a bit of preliminary work. We may view the essential requirement one of synthesizing sentences. This activity may be compared to synthesizing protein molecules--though nothing like the expenditure of time and money has been applied to linguistic investigation as to that of chemistry.

One of the first problems we may note is that language is not a simple linear structure; rather, it consists of numerous structures. One is made up of entities like t m and so on, which might be compared with atoms; this structure contains relatively few entities, but their rules of interrelationship are complex. A second structure is made up of entities like then men walk, which might be compared with radicals; this structure contains a great number of entities, possibly with somewhat less complex rules of interrelationships. From these the smallest free form of language is constructed, the sentence. In making sentences, in using language man has somehow learned to master both of these structures. More, he has learned the relationships of the entities in the second structure to a totally different structure, that of concepts. Since computer manipulation of language is a type of simulation, before we can use computers effectively for managing language, we must understand how these various structures relate to one another, how language functions.

Of the various problems, some are straightforward, for example, the amassing of entities and their rules. We spend the first ten years of our lives acquiring control over one language, continue to add to our stock of entities, and rarely achieve mastery over a second language. To give the computer similar opportunities we must have large-scale programs, by means of which we can store genuine materials and materials with a characteristic meaning. A great deal of effort has been expended by members of the Linguistics Research Center over the past five years to develop the system of programs which handle the data of language and their interrelationships.

Man has taken care of this problem very cleverly. He reduces language to a set of entities of sound, about forty in a language, and accordingly has relatively few building blocks to control. Unfortunately no machine has been devised to match man's discriminatory powers in managing the entities of speech. Accordingly at present, machine manipulation of language must be based on the second level with its tremendous number of entities.

Since our work in the Center is still experimental, it is difficult to forecast how many such entities must be stored in a computer. Some estimates put the number of chemical terms in German at two million; the rules for relating them to other entities of the language will obviously be fewer.

In the relatively small computers available today the rules indicating interrelationships and vocabulary items of a language must be stored on magnetic tape. The programming system developed at the Linguistics Research Center has been successful in analyzing materials of limited vocabulary and syntactic complexity, consisting of about 50,000 rules and items in each language.

Until one has dealt with the highly rigorous computer it is almost impossible to visualize the problems involved in a thorough analysis of language. A simple example from a physics textbook may illustrate some of them:

> Loudness is the property of sound determined by the effect of the power of the sound waves on our ears.

Let us suppose that we write a computer routine--a syntactic rule--relating <u>of</u> to a following noun, as in <u>of sound</u>; the rule will not then handle the sequence <u>of the power</u>, for here <u>of</u> is followed by the definite article. If we modify our rule accordingly, we still have not handled the use of <u>of</u> in <u>of the sound waves</u>, for here the article is followed by a noun used as adjective. In putting a sentence like this into a computer, we must therefore provide for sequences of <u>of</u> and a variety of entities. Obviously our rules cannot be simple, though our example may have been.

Another entity of the sentence, <u>on</u> may illustrate a further type of problem. If we relate <u>on</u> to the surrounding entities, we arrive at the possible sequence: <u>the sound waves on our ears</u>. This sequence might compare to that of <u>the flag waves on the</u>

flag-pole or the policeman waves on the traffic. But such relationships which would appear identical to a computer trouble us; the sentence from our physics textbook seems absurd to us if on is related to waves in either of these ways. We have learned that on is related to effect and that the meaningful sequence is effect on our ears. We scarcely need to discuss the inadequate translations that would be produced if such sentences were put word-for-word into German, Russian or other languages.

Since an English speaker understands his language in this way, a computer must also be prepared to manipulate it accordingly. To arrange such manipulations we must describe English far more precisely than has ever been done before. The required detail of description has never been provided before because native speakers master such sequences, and we are charitable to foreigners who learn inadequate English from our inadequate grammars. But if a computer makes any requirement, it is for precision. A computer would not be happy with our simple sentence until it knows what to do with every entity, including on. Consequently a linguist has to determine the role of an entity first of all, then describe it. Since even the large dictionaries which have been produced for English, German and the other widely studied languages have not described these languages adequately, linguists in the Linguistics Research Center are now at work producing such descriptions--writing rules for English, German, Russian and other languages. Figures 1 - 5 illustrate the procedures involved in making a syntactic analysis of an English sentence, in accordance with a grammar written by Dr. Wayne Tosh of our Center.

The resultant rules are many and intricate. When produced, they must be handled by the computer, but kept independent of it through use of generalized computer programming. If, for example, a specialized program were written for handling combinations of prepositions plus nouns or prepositions plus articles plus nouns, it would have to be revised to handle sequences of prepositions plus articles plus noun-adjectives plus nouns, as in <u>of the sound waves</u>. The Linguistic Research System, produced under the direction of Eugene Pendergraft, was devised to meet this requirement of generalization. With this system linguistic rules, independent of specialized computer programs, may be produced to handle phrases of various length--preposition plus noun as in <u>of sound</u>, preposition plus article plus noun, as in <u>of the power</u>, and longer phrases like <u>of the sound waves</u>. Other instructions alert the computer to watch out for prepositions like <u>on</u> after a noun like <u>effect</u>. A chart of the system indicates the demands placed on computer manipulation of language and also one of the results of five years of work, supported by the US Army Electronics Laboratories and by the National Science Foundation.

One of our practical problems is to achieve an understanding by outsiders of the use of computers in processing linguistic material. Most of us had our notions about scientific procedures determined by elementary science class, in which we probably used a Bunsen burner. This early activity seems to leave the indelible impression that scientific equipment, for example a computer, is like a Bunsen burner. There is little variety of use for a Bunsen burner--it merely heats things. The heat isn't different if one lights it with a flint, or a match--if one strikes the match on a piece of sandpaper or one's thumbnail. By analogy it is assumed that the machine is the essential part of computation; after one switches on the power, a computer can cook your data as well as mine. Yet in language, as in the social sciences, the important part of computation is the program. The importance of how one utilizes a machine rather than the makeup of the machine may be one of the essential differences between work in the social sciences and that in the natural sciences. Possibly software and hardware sciences would be more appropriate names.

5 AUTOMATIC CLASSIFICATION

Yet even a system with programs of the complexity of those illustrated is inadequate for handling language. In a sentence of fewer than twenty elements, for example, there are more than a million possibilities of analysis. But this figure, large though it is, fails to take into account an analysis for meaning, for determining among other things that in our sentence sound is similar in meaning to noise rather than to healthy, valid, as in a sound mind or a sound theory. When we handle the multitude of entities necessary in analysis of meaning we will deal with many more possibilities of interpretation than are found for on. In managing these, our present computers would be choked. Even the larger computers now becoming available would deal with the quantities of data slowly. Adequate speed seems possible only by refinements of computer theory and in improved techniques of classification.

A few years ago R. M. Needham, of Cambridge University, pointed the way to such classification with his clumping theory. His procedures are being expanded for application to larger sets of data by A. G. and N. Dale of our Center. Details are provided in the paper, A Programming System for Automatic Classification with Applications in Linguistic and Information Retrieval Research, LRC 64 WTM-4, written by A. G. and N. Dale and E. D. Pendergraft. With other papers, this is available from the Center. Even the procedures described in this paper require a great deal of computer time for handling a relatively small number of entities. Further research is being pursued to improve and speed up the procedure. I have time merely to mention such research, but would also like to point out that it was not even envisaged before language analysis with computers was undertaken. The amount of linguistic data which must be manipulated, as well as its complexity, has pointed up the need for research in fields of applied logic or mathematics that would not have been related to language investigation a few years ago. Students in the sciences, for whom the required language courses may seem to have little lasting value, might well consider applying themselves to these problems. Solutions will follow only from a quantitative approach, generally lacking in previous students of language.

6 ANALYSIS OF MEANING

But though we face numerous problems in the development of computer systems and in the theoretical work which must be carried out before systems and computers can manage efficiently the huge and complex amounts of data, our largest problems remain in the understanding of language. Chief among these is the treatment of meaning. In dealing with meaning we are probably a bit farther along than Plato, though not much. Our dictionaries largely side-step the problem; they set out to provide synonyms, whether monolingually or bilingually. Since they are fairly effective tools, we can handle translation of a sort without understanding meaning. But for competent translation, for automating indexing and abstracting, for problems in artificial intelligence, we will have to control meaning as we now do syntactic relationships.

Our theoretical approach is clear. We assume that language is structured at the level of meaning similarly to its structure at the levels of sound and syntax. Again, we do not relate entities to the outside world, but to concepts. Still the problems of analysis are staggering. The sheer magnitude of the data--all human knowledge--is troublesome enough. But how to classify it? By specialties as we do in real life? Should one computer handle nuclear physics, another the physics of light, another molecular biology, and so on? (If we did, we would not welcome a physicist who also concerns himself with biology). But if we divide the universe of concepts in this way, what type of hierarchical arrangement should we use? If, for example, we define _man_ as 'male human being', should we distinguish between the concepts 'male' and 'human being' because 'male' is automatically supplied in such sequences as 'he was a man who..., the king is a man who...'? It will be difficult to answer such questions until we carry on a fair bit of investigation. Before then, it will even be difficult to pose the proper questions.

7 ACCOMPLISHMENTS

It may be disappointing for non-linguists to hear that linguistic work has scarcely begun, with or without computers. Be we have some accomplishments. Some theoretical positions seem supportable. We are on our way to an extensive and flexible linguistic research system, and expect to have adequate computers to make use of it. The traditionally lone linguist is beginning to work with specialists in related fields. Even the achievement of analyzing language syntactically may seem small. But our tools are still inadequate. Given satisfactory scanning devices and more powerful computers we will be able to use our system for analyzing more than a snatch of language. Already straightforward linguistic applications may be carried out, if adequate resources are provided; any book may be automatically indexed, and accordingly among other things more readily proof-read. Bibliographical and other data may be managed automatically; in a pilot project, the Center has listed all Slavic books in the University Library, so that anyone interested in Tolstoy, in Russian novels or the like, may be given an immediate print-out of the titles. Other such projects need only financial support for achievement. The chief aim of the Center, however, is to continue theoretical investigations of language and data processing techniques, and the preparation of computer programs, so that ultimately a computer will be able to manipulate language with somewhat the same proficiency as does man.

1. INPUT SENTENCE
TO
ANALYSIS PROGRAM

SCIENCE CONFERENCE CORPUS
CORPUS DISPLAY

20 NUVEMBER 1964

```
01CCC001       UNIVERSITY OF TEXAS SCIENCE CONFERENCE
C1COC002       NOVEMBER 20, 1964
01000003
01001001  THIS IS A SENTENCE ANALYZED BY THE LINGUISTICS RESEARCH SYSTEM.
01CC1C02
C1002001 (SEE ACCOMPANYING DISPLAYS)
01002002
```

```
SCIENCE CONFERENCE CORPUS                          20 NOVEMBER 1964
CORPUS DISPLAY

PAGE 1
                   ..........10‡.........20‡.........30‡.........40‡.........50‡.........60‡.........70‡.........80‡.........90‡.........100‡
                              ‡         ‡         ‡         ‡         ‡         ‡         ‡         ‡         ‡         ‡
  01   001    000  THIS IS A‡ SENTENCE ‡ANALYZED B‡Y THE LING‡UISTICS RE‡SEARCH SYS‡TEM.      ‡         ‡         ‡         ‡
```

2. MATRIX IDENTIFYING CHARACTER POSITIONS IN INPUT SENTENCE

Sentence begins in col. 2 and ends in col. 64.

SCIENCE CONFERENCE GRAMMAR

20 NOVEMBER 1964

SYNTACTIC DATA----STRATUM 2--FORM SORT

NOTES	FORM			DESIGNATUM				
D 0 P 1.000000 -0	2	C	67	1	V DTRMNR 1S	* THIS		
D 0 P 1.000000 -1	2	C	72	1	V N5A	* SEN	* TENCE B	
D 0 P 1.000000 -0	2	C	75	1	V V4C	* ANALYZ		
D 0 P 1.000000 -0	2	C	77	1	V PRPSTN	* BY		
D 0 P 1.000000 -0	2	C	79	1	V DTRMNR 1	* THE		
D 0 P 1.000000 -0	2	C	82	1	V N3A	* LIN	* GUIS B	* TICS B
D 0 P 1.000000 -0	2	C	84	1	V N5H	* RE	* SEARCH B	
D 0 P 1.000000 -1	2	C	85	1	V N5A	* SYS	* TEM B	

3. GRAMMAR USED IN AUTOMATIC ANALYSIS

Symbols in DESIGNATUM column are class names of construction lying to right of symbol. Each entry is separate rule identified uniquely by number in FORM column.

SYNTACTIC DATA----STRATUM 3--FORM SORT

```
NOTES          FORM          DESIGNATUM

D 1            2  C  65      1   V SNTNC    V CLS      * .
P 1.000000                                  S 1        B
   -0

D 2            2  C  66      1   V CLS      V DTRMNR   V BE
P 1.000000                                    1S         SNGLR
   -0                                       S 1          PRSNT
                                                       S 2

D 1            2  C  68      1   V BE       * IS       V NP
P 1.000000                         SNGLR                 SNGLR
   -0                              PRSNT               S 1

D 1            2  C  69      1   V NP       * A        V NMNL
P 1.000000                         SNGLR                 A
   -1                                                  S 1

D 2            2  C  70      1   V NMNL     V NMNL     V VRBL
P 1.333312                         A          A          -D2
   -2                                         /6/S       PHRS
                                            S 1        S 2

D 1            2  C  71      1   V NMNL     V N5A
P 1.000000                         A        S 1
   -1                              /6/S

D 2            2  C  73      1   V VRBL     V VRBL     V PRPSTN
P 1.000000                         -D2        -D2AN      PHRS
   -0                              PHRS     S 1        S 2

D 1            2  C  74      1   V VRBL     V V4C      * ED
P 1.000000                         -D2AN    S 1        B
   -0

D 2            2  C  76      1   V PRPSTN   V PRPSTN   V NP
P 1.000000                         PHRS     S 1          SNGLR
   -0                                                  S 2
```

```
NOTES          FORM            DESIGNATUM

D 2            2  C 78         1  V NP      V DTRMNR   V NMNL
P 1.000000                          SNGLR      1          A
   -1                                       S  1       S  2

D 2            2  C 80         1  V NMNL    V NMNL     V N5A
P 1.333312                           A         A       S  2
   -2                                       S  1

D 2            2  C 81         1  V NMNL    V N3A      V NMNL
P 1.333312                           A      S  1          A
   -2                                                     /6/S
                                                       S  2

D 1            2  C 83         1  V NMNL    V N5H
P 1.000000                           A      S  1
   -1                                /6/S
```

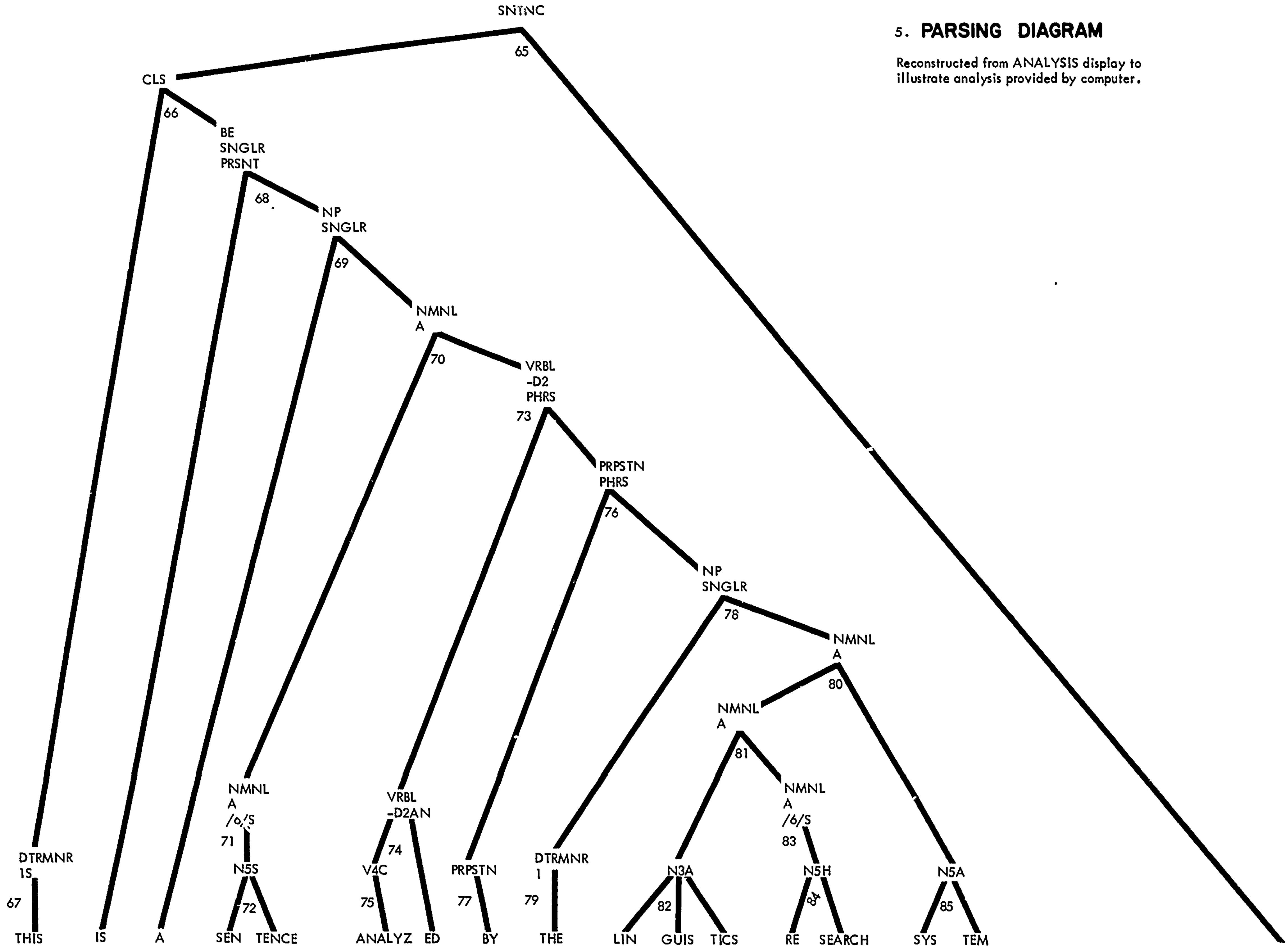
FROM	TO	PROBABILITY													NOTES
10	10		*A												
10	11		A												
10	56	1.77770-9		69 82	70 83	71 84	72	73	74	75	76	77	78	79	81
10	63	1.18511-11		69 81	80 82	70 83	71 84	72 85	73	74	75	76	77	78	79
10	63	1.18511-11		69 81	70 82	71 83	72 84	73 85	74	75	76	77	78	79	80
11	11		*												
12	14		*SEN												
12	19	1.00000-1		72											
12	19	1.00000-2		71	72										
12	20		SENTENCE												
12	56	1.77770-8		70 83	71 84	72	73	74	75	76	77	78	79	81	82
12	63	1.18511-10		80 82	70 83	71 84	72 85	73	74	75	76	77	78	79	81
12	63	1.18511-10		70 82	71 83	72 84	73 85	74	75	76	77	78	79	80	81
15	19		*TENCE												
20	20		*												
20	20														
21	21		*A												
21	26		*ANALYZ												
21	26	1.0		75											
21	28	1.0		74	75										
21	56	1.33331-4		73	74	75	76	77	78	79	81	82	83	84	
21	63	1.77770-7		73 85	74	75	76	77	78	79	80	81	82	83	84

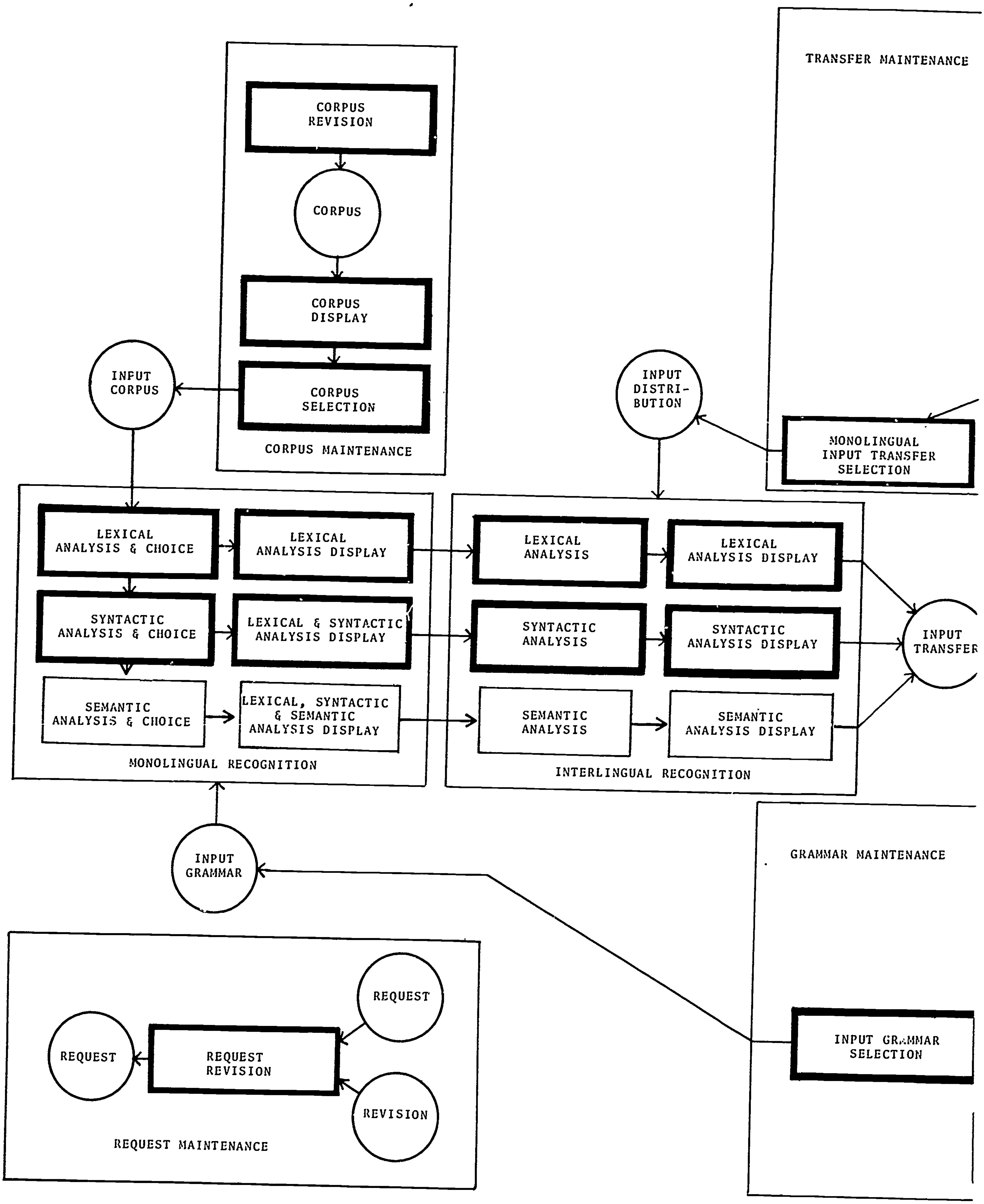
CORPUS 01 SAMPLE 001

FROM	TO	PROBABILITY												NOTES
23	23		*A											
27	28		*ED											
27	28		ED											
27	29		ED											
29	29		*											
29	29													
30	31		*BY											
30	31	1.0		77										
30	32		BY											
30	56	1.33331-4		76	77	78	79	81	82	83	84			
30	63	1.77770-7		76	77	78	79	80	81	82	83	84	85	
32	32		*											
32	32													
33	35		*THE											
33	35	1.0		79										
33	36		THE											
33	56	1.33331-4		78	79	81	82	83	84					
33	63	1.77770-7		78	79	80	81	82	83	84	85			
36	36		*											
36	36													
37	39		*LIN											
37	47	1.0		82										
37	48		LINGUISTICS											
37	56	1.33331-3		81	82	83	84							
37	63	1.77770-6		80	81	82	83	84	85					

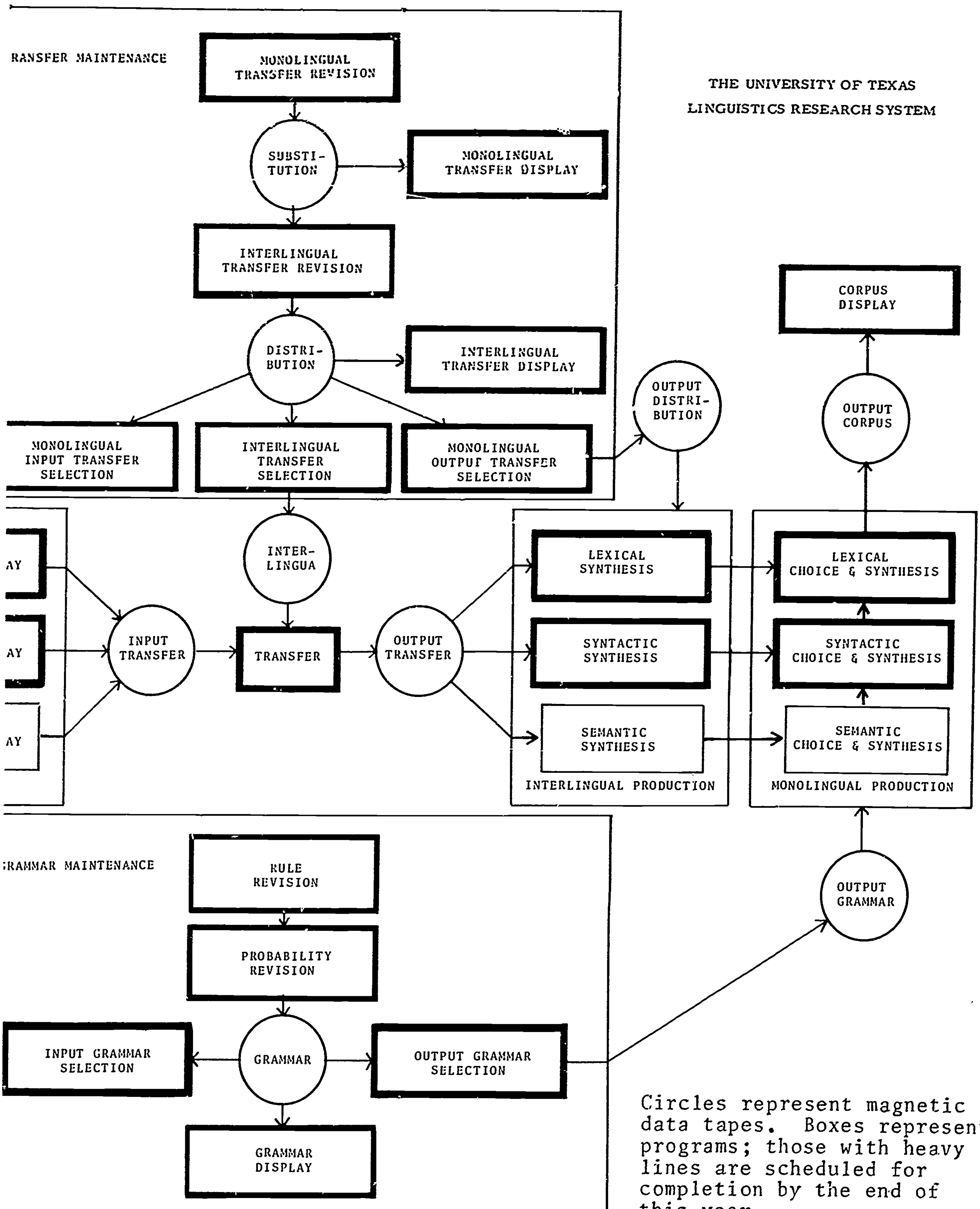
5. PARSING DIAGRAM

Reconstructed from ANALYSIS display to illustrate analysis provided by computer.




TRANSFER MAINTENANCE
CORPUS REVISION
CORPUS
CORPUS DISPLAY
CORPUS SELECTION
CORPUS MAINTENANCE
INPUT CORPUS
INPUT DISTRI- BUTION
MONOLINGUAL INPUT TRANSFER SELECTION
LEXICAL ANALYSIS & CHOICE
LEXICAL ANALYSIS DISPLAY
SYNTACTIC ANALYSIS & CHOICE
LEXICAL & SYNTACTIC ANALYSIS DISPLAY
SEMANTIC ANALYSIS & CHOICE
LEXICAL, SYNTACTIC & SEMANTIC ANALYSIS DISPLAY
MONOLINGUAL RECOGNITION
LEXICAL ANALYSIS
LEXICAL ANALYSIS DISPLAY
SYNTACTIC ANALYSIS
SYNTACTIC ANALYSIS DISPLAY
SEMANTIC ANALYSIS
SEMANTIC ANALYSIS DISPLAY
INTERLINGUAL RECOGNITION
INPUT TRANSFER
INPUT GRAMMAR
GRAMMAR MAINTENANCE
INPUT GRAMMAR SELECTION
REQUEST
REQUEST
REQUEST REVISION
REVISION
REQUEST MAINTENANCE

THE UNIVERSITY OF TEXAS
LINGUISTICS RESEARCH SYSTEM



Circles represent magnetic data tapes. Boxes represent programs; those with heavy lines are scheduled for completion by the end of this year.